README

Documentación Técnica — Proyecto UT1 · Retail Mini (Batch Simple)

Esta carpeta contiene toda la documentación técnica del proyecto **Retail Mini**, correspondiente a la Unidad de Trabajo 1 (UT1).

El proyecto implementa un pipeline de tratamiento de datos con el siguiente flujo:

CSV (Bronce) → Limpieza y modelado (Plata) → Reporte final (Oro)

Se basa en un caso de negocio de tipo *Retail Mini*, con ventas diarias y estructura batch simple.

Diseño de la Ingesta

10. Diseño de la Ingesta de Datos

Definición del proceso de ingesta de datos de ventas diarias en formato CSV, garantizando idempotencia, trazabilidad y control de calidad.

El proceso de ingesta se ejecuta mediante *batch diario* desde ficheros CSV ubicados en la ruta `project/data/drops/YYYY-MM-DD/ventas.csv`.

Cada lote se identifica con un `batch_id` único y conserva los metadatos `_source_file`, `_ingest_ts` y `_batch_id` para trazabilidad.

La política de deduplicación aplica “último gana” por la clave natural (`fecha`, `id_cliente`, `id_producto`).

El sistema valida estructura, tipos y valores antes de almacenar los datos. Los registros inválidos se guardan en `project/output/quality/ventas_invalidas.csv`.

Limpieza y Calidad

20. Limpieza y Control de Calidad

El proceso de limpieza garantiza que los datos sean consistentes antes de modelarlos.

1. Conversión de tipos de datos (`fecha`, `unidades`, `precio_unitario`).

2. Validación de rangos (`unidades >= 0`, `precio_unitario >= 0`).

3. Detección de nulos o vacíos en identificadores.

4. Registros válidos → `clean_ventas.parquet`

5. Registros inválidos → `ventas_invalidas.csv`

Se aplica deduplicación sobre la clave natural y control de integridad referencial.

Se registran métricas de calidad y porcentaje de registros válidos por lote.

Modelado de Datos

30. Modelado de Datos y KPIs

El modelado se realiza en la capa oro. Se genera el campo derivado `importe = unidades * precio_unitario`.

Los datos limpios se guardan en:

- Parquet: `clean_ventas.parquet`

- SQLite: `ut1.db` (tablas `raw_ventas`, `clean_ventas`)

Principales KPIs:

- Ingresos totales

- Ticket medio

- Transacciones

- Producto líder

- Cobertura temporal

Se crean vistas SQL como `ventas_diarias` para análisis agregados.

La moneda base es el euro (EUR) y los nulos se tratan como 0.

Reporte Final

40. Reporte Markdown — Resultados

El reporte Markdown integra todos los indicadores y tablas resultantes del pipeline.

Secciones:

1. Titular del periodo

2. KPIs principales (ingresos, ticket medio, transacciones)

3. Top productos

4. Resumen diario

5. Calidad y cobertura

6. Conclusiones y recomendaciones

Ejemplo:

Periodo: 2025-01-01 a 2025-01-31

Ingresos totales: 12 340.50 €

Ticket medio: 17.34 €

Transacciones: 712

Producto líder: P10

Lecciones Aprendidas

99. Lecciones Aprendidas

Aspectos positivos

- Modularidad del pipeline (ingesta, limpieza, modelado, reporte)

- Uso eficiente de Parquet y SQLite

- Idempotencia y trazabilidad reproducibles

Dificultades

- Configuración de Quartz y GitHub Pages

- Ajuste de versiones Node y dependencias

- Validación de formatos mixtos

Mejoras futuras

- Automatización de calidad

- UI para selección de lotes

- Alertas ante errores

Conclusión

El proyecto cumple los objetivos de la UT1 y constituye una base sólida para flujos ETL más complejos.